

Location Prediction: Communities Speak Louder than Friends

Jun Pang
University of Luxembourg
FSTC & SnT
jun.pang@uni.lu

Yang Zhang
University of Luxembourg
FSTC
yang.zhang@uni.lu

ABSTRACT

Humans are social animals, they interact with different communities of friends to conduct different activities. The literature shows that human mobility is constrained by their social relations. In this paper, we investigate the social impact of a person's communities on his mobility, instead of all friends from his online social networks. This study can be particularly useful, as certain social behaviors are influenced by specific communities but not all friends. To achieve our goal, we first develop a measure to characterize a person's social diversity, which we term 'community entropy'. Through analysis of two real-life datasets, we demonstrate that a person's mobility is influenced only by a small fraction of his communities and the influence depends on the social contexts of the communities. We then exploit machine learning techniques to predict users' future movement based on their communities' information. Extensive experiments demonstrate the prediction's effectiveness.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—*Data mining*

General Terms

Algorithms, theory, experiments

Keywords

Human mobility, social networks, network communities

1. INTRODUCTION

Humans are social animals, everyone is a part of the society and gets influences from it. For example, our daily behaviors, such as what types of music we listen to, where we have lunch on weekdays and what activities we conduct on weekends, are largely dependent on our social relations. Normally, we categorize our social relations into different

groups, i.e., social communities, using different criteria and considerations. By definition, *a community is a social unit of any size that shares common values*.¹ Typical communities include family, close friends, colleagues, etc. In daily life, humans are engaged in various social environments, and they interact with different communities depending on the environments. For our specific behaviors, social influences, in most of cases, are not from *all our friends* but from *certain communities*. For example, we listen to similar types of music as our close friends, but not as our parents; we have lunch together with our colleagues on weekdays, but not with our college friends living in another city; on weekends we spend more time with family, but not with our colleagues.

Location-based social network services (LBSNs) have been booming during the past five years. Nowadays, it is common for a user to attach his location when he publishes a photo or a status using his online social network account. Moreover, users may just share their locations, called *check-in*, to tell their friends where they are or to engage in social games as in Foursquare. Since these large amount of location and social relation data become available, studying human mobility and its connection with social relationships becomes quantitatively achievable (e.g., [16, 10, 9, 34, 15, 7, 8]). Understanding human mobility can lead to compelling applications including location recommendation [44, 41, 43, 14, 20], urban planning [42], immigration patterns [5], etc.

Previous works, including [1, 6, 10, 33], show that human mobility is influenced by social factors. However, there is one common shortcoming: they all treat friends of users equally. Similar to other social behaviors, in most cases mobility is influenced by specific communities but not all friends. For example, the aforementioned colleagues can influence the place a user goes for lunch but probably have nothing to do with his weekend plans. Meanwhile, where a user visits on weekends largely depends on his friends or family, but not his colleagues. Therefore, the impact on a user's mobility should be considered from the perspectives of communities instead of all friends. In a broader view, community is arguably the most useful resolution to study social networks [39].

Contributions. In this paper, we aim to study the impact from communities on a user's mobility and predict his locations based on his community information.

First, we partition each users' friends into communities and propose a notion namely community entropy to quantify a user's social diversity. Second, we analyze communities' influences on users' mobility and our main conclusions include:

¹<http://en.wikipedia.org/wiki/Community>

(1) communities’ influences on users’ mobility are stronger than their friends’; (2) each user is only influenced by a small number of his communities; and (3) such influence is typically constrained by temporal and spatial contexts. Third, we predict users’ locations using their community information. Experimental results on two real-life datasets with millions of location data show that the community-based predictor achieves a strong performance.

Organization. After the introduction, we present a few preliminaries and our datasets in Section 2. Then we describe the community detection process and propose the notion of community entropy in Section 3. The relationship between users and their communities on mobility is analyzed in Section 4. Based on our analysis, we propose a location predictor with features linked to community information and present experimental results in Section 5. We discuss related work in Section 6 and conclude our paper with some future work in Section 7.

2. PRELIMINARIES

We summarize the notations in Section 2.1 and describe the datasets that we use throughout the paper in Section 2.2.

2.1 Notations

All users are contained in the set \mathcal{U} while a single user is denoted by u . We use the set $f(u)$ to represent u ’s friends. A community of a user u is a subset of his friends denoted by c and $c \subseteq f(u)$. Meanwhile, $C(u)$ represents all the communities of u , i.e., $C(u)$ is a set of sets of u ’s communities. Every friend of a user is assigned into one of the user’s communities, the union of all his communities is the set of all his friends. In this work, we only consider non-overlapping communities, namely $c \cap c' = \emptyset$ for $c, c' \in C(u)$. However, this assumption is not crucial to our approach and our results can be extended for overlapping communities as well.

A check-in of u is denoted by a tuple $\langle u, t, \ell \rangle$, where t represents the time and ℓ is the location that corresponds to a pair of latitude and longitude. We use $CI(u)$ to represent all the check-ins of u . Without ambiguity, we use location and check-in interchangeably in the following discussion.

2.2 The datasets

We exploit two types of social network datasets for this work. The first one is collected by the authors of [10] from Gowalla – a popular LBSN service back in 2011. The dataset was collected from February 2009 to October 2010 and it contains 6,442,892 check-ins. Besides location information, the dataset also includes the corresponding social data which contains around 1.9 million users and 9.5 million edges. Due to the large data sparsity, we mainly focus on the check-in data in two cities in US, including New York (NY (G)) and San Francisco (SF (G)). They are among the areas with most check-ins in the dataset. In addition, when performing mobility analysis and location prediction, we only focus on users who have conducted at least 100 check-ins in each city and we term these users as *active users*.

The second dataset is collected from Twitter from December 2014 to April 2015 by the authors of this paper. Again, we focus on the data in New York (NY (T)) and San Francisco (SF (T)) and treat all the geo-tagged tweets (tweets labeled with geographical coordinates) as users’ check-ins.

We exploit Twitter’s Streaming API² to collect all the geo-tagged tweets. Each check-in is organized as a 4-tuple.

$$\langle uid, time, latitude, longitude \rangle$$

Figure 1 depicts a sample of check-ins in New York. To collect the social relationships among users, we adopt Twitter’s REST API³ to query each user’s followers and followees. Two users are considered friends if they follow each other mutually.

Similar to the Gowalla dataset, we only focus on active users (users with more than 100 check-ins) in the Twitter dataset. Moreover, we also filter out the users who have more than 2,000 check-ins since most of them are public accounts such as @NewYorkCP which publishes 16,681 check-ins at the exact same location. Table 1 summarizes the two datasets. The Twitter dataset is available upon request.

3. COMMUNITIES

We first show how to detect communities in social networks in Section 3.1 and then propose a new notion to characterize users’ social diversity in Section 3.2.

3.1 Community detection in social networks

Community detection in networks (or graphs) has been extensively studied for the past decade (e.g., see [25, 31, 2, 18, 32, 22, 38, 37, 21, 39, 23]). It has important applications in many fields, including physics, biology, sociology as well as computer science. The principle behind community detection is to partition nodes of a large graph into groups following certain metrics on the graph structure [18]. In the context of social networks, besides the social graph, each user is also affiliated with attributes. These information can also be used to detect communities (e.g., see [22, 38, 23]). For example, people who graduate from the same university can be considered as a community. Since the datasets we use only contain social graphs and no personal information are provided, we apply the algorithms that are based on information encoded in graph structure to detect communities.

According to the comparative analysis [18], among all the community detection algorithms, Infomap [31] has the best performance on undirected and unweighted graphs and has been widely used in many systems [26, 29]. Therefore, we apply it in this work. Next we give a brief overview of Infomap and describe how we use it to detect communities.

The main idea of Infomap can be summarized as follows: information flow in a network can characterize the behavior of the whole network, which consequently reflects the structure of the network. A group of nodes among which information flows relatively fast can be considered as one community. Therefore, Infomap intends to use information flow to detect communities in a network. In the beginning, Infomap simulates information flow in a network with random walks. Then the algorithm partitions the network into communities and exploits Huffman coding to encode the network at two levels. At the community level, the algorithm assigns a unique code for each community based on the information flow among different communities; at the node level, the algorithm assigns a code for each node based on the information flow within the community. Infomap allows the Huffman codes in different communities (node level) being

²<https://dev.twitter.com/streaming/overview>

³<https://dev.twitter.com/rest/public>

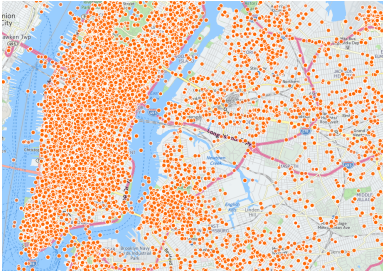


Figure 1: Check-ins in New York.

| | NY (G) | SF (G) | NY (T) | SF (T) |
|--------------------------------|---------|---------|-----------|-----------|
| # of users | 7,786 | 6,617 | 207,805 | 113,383 |
| # of check-ins | 176,324 | 177,357 | 2,325,907 | 2,163,959 |
| Avg.# of check-ins | 21.6 | 26.8 | 11.2 | 19.1 |
| # of active users | 175 | 236 | 1,636 | 1,626 |
| Avg.# of friends (active user) | 79.4 | 69.7 | 376.9 | 289.0 |

Table 1: Summary of the datasets.

uplicated which results in a more efficient encoding (less description length). In the end, finding a Huffman code to concisely describe the information flow while minimizing the description length is thus equivalent to discovering the network’s community structure. In other words, the objective of Infomap is to find a partition of a network such that the code length for representing information flow among communities and within each community is minimized. Since it is infeasible to search all possible community partitions, Infomap further exploits a deterministic greedy search algorithm [11, 36] to find partitions.

In our work, to detect communities of u , we first find all his friends as well as the links among them. Then, we delete u and all edges linked to him and apply Infomap algorithm to the remaining part of the graph. Figure 2 presents the detected communities of two users in the Gowalla dataset. Each community is marked with a different color.

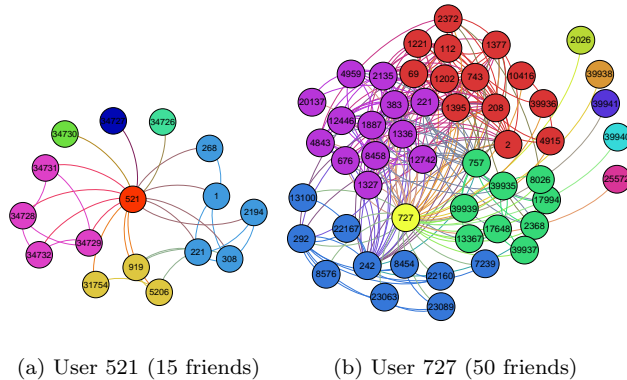


Figure 2: Communities of users 521 and 727.

| | Gowalla | Twitter |
|----------------------|---------|---------|
| Avg.# of communities | 4.5 | 5.3 |
| Avg. community size | 13.2 | 20.8 |

Table 2: Community summary of active users.

Table 2 lists the summary of community information of all active users in the two datasets. Each active user in Gowalla has on average 4.5 communities while the value is 5.3 for the Twitter users. In addition, the average community size of Twitter users is bigger than Gowalla users (20.8 vs. 13.2). This is because active users in the Twitter dataset have more friends than those in the Gowalla dataset

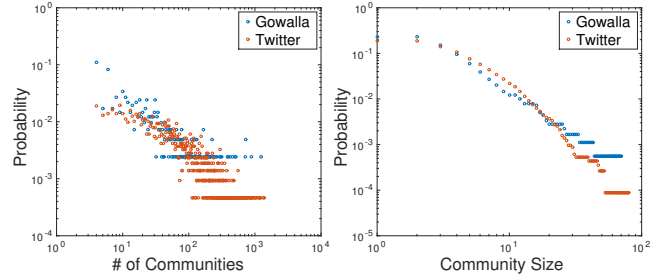


Figure 3: Distribution of users w.r.t the number of communities and distribution of communities w.r.t their size.

(see Table 1), which indicates general social network services, such as Twitter, contain more users’ social relationships than LBSN services, such as Gowalla. In spite of the differences on the average value in Table 2, community number and community size in the two datasets follow a similar distribution. As we can see from Figure 3, both community number and size follow the power law: most of the users have small number of communities and most of the detected communities are small as well.

3.2 Community entropy

After detecting communities, we are given a new domain of attributes on users. We are particularly interested in how diverse a user’s communities are. We motivate this *social diversity* through an example. Suppose that a user is engaged in many communities, such as colleagues at work, family members, college friends, chess club, basketball team, etc, then he is considered an active society member. Users of this kind are always involving in different social scenarios or environments, and his daily behaviors are largely dependent on his social relations.

Although we do not have the semantics of each of our detected communities, such as the aforementioned colleagues at work or chess club, we can still use the information encoded in the graph to define a user’s social diversity. For instance, for a user with several communities whose sizes are more or less the same, his social diversity is for sure higher than those with only one community.

To quantify the social diversity of a user, we introduce the notion of *community entropy*.

Definition 1. For a user u , his *community entropy* is defined as

$$coment(u) = \frac{1}{1-\alpha} \ln \sum_{c \in C(u)} \left(\frac{|c|}{|f(u)|} \right)^\alpha.$$

Our community entropy follows the definition of Rényi entropy [30]. Here, α is called the order of diversity, it can control the impact of community size on the value which gives more flexibility to distinguish users when focusing on the sizes of their communities. In simple terms, our community entropy,

- when $\alpha > 1$, values more on larger communities;
- when $\alpha < 1$, values more on smaller communities.

The limit of $coment(u)$ with $\alpha \rightarrow 1$ is the Shannon entropy.⁴ In general, if a user has many communities with sizes equally distributed, then his community entropy is high and this indicates that his social relations are highly diverse.

We set $\alpha > 1$ in the following discussion to limit the impact of small communities since a user may randomly add strangers as his friends in online social networks and these strangers normally form small communities (such as a one-user community⁵), which have less impact on the user’s mobility. For example, if a user u has three communities with sizes equal to 1, 1 and 10, then his communities are not that diverse following the above intuition. When we set α less than 1, such as 0.5, we have $coment(u) = 0.79$ which is a high value indicating u ’s social circles are diverse. On the other hand, if we set α bigger than 1, such as 10, $coment(u)$ drops to 0.20 which captures our intuition. In the following experiments, we set $\alpha = 10$ when calculating users’ community entropies. Note that we have also set α to other numbers bigger than one and observed similar results. Figure 4 shows the histogram of community entropies of all active users in two datasets.

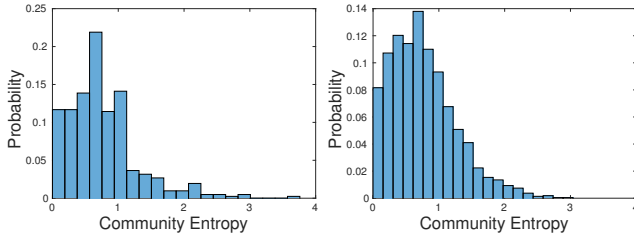


Figure 4: Distribution of community entropies of active users in Gowalla (left) and Twitter (right).

4. COMMUNITIES AND MOBILITY

It has been proved that social factors play an important role on users’ mobility, e.g., see [10]. For instance, one may go to lunch with his friends or go to a bar to hangout with his friends. Meanwhile, for a user, friends of his social networks (as well as in real life) are not all equal. Instead friends normally belong to certain communities. When considering a user’s mobility, intuitively different communities can impose different influence within certain contexts or social environments. Continuing with the above example, the people the user has lunch with are normally his colleagues while the people he meets at night are his close friends. Therefore, in order to analyze the impact from a user’s social relations on

⁴https://en.wikipedia.org/wiki/Renyi_entropy

⁵In our community detection algorithm, if u ’ himself forms a community of u , then it indicates that u ’ does not know any other friends of u .

his mobility, it is reasonable to focus on social influence at the community level.

In this section, we first study communities’ influence on users’ mobility. After that, we study the characteristics of the influential communities with the following two intuitions in mind: (1) a user’s daily activities are constrained, and the number of communities he interacts with is limited; (2) communities influence a user’s social behavior under different contexts.

4.1 Influential communities

Figure 5a depicts a user’s two communities’ check-ins in Manhattan of the New York City. We can observe a quite clear separation between these two communities’ check-ins: members of community 1 mainly visit Uptown and Midtown Manhattan while community 2 focuses more on Midtown. This indicates that different communities have their social activities at different areas. In a broader view, this shows that partitioning users’ check-ins at the social network level (through community detection) can result in meaningful spatial clusters as well.

A single community also has several favorite places. For example, community 1 in Figure 5a visits Times Square and Broadway quite often while members of community 2 like to stay close to Madison square park. A user may socialize with different communities at different places, for example, he may go to watch a basketball game with his family at the stadium and have lunch with his colleagues near his office. Therefore, to study influences on mobility from communities to a user, we need to summarize each community’s *frequent movement areas*. To discover a community’s frequent movement areas, we perform clustering on all locations that the community members have been to. Each cluster is then represented by its central point and a community’s frequent movement areas are thus represented by the centroids of all clusters. The clustering algorithm we use is the agglomerative hierarchical clustering. We regulate that any two clusters can be aligned only if the distance between their corresponding centroids is less than 500m which is a reasonable range for human mobility.

To illustrate the mobility influence from communities to users, we choose to use ‘distances’. More precisely, we represent the influence by the distances between a user’s locations and the frequent movement areas of his communities. Shorter distances imply stronger influences. For each location a user has visited, we calculate the distances between the location and all his communities’ frequent movement areas. Then, for each community of the user, we choose the shortest distance between the location and the community’s frequent movement areas as the *distance* between the location and the community. The community which has the smallest distance to the location is considered as the *influential community* of the user at this location. The distance between the influential community and the user’s location is further defined as the distance between the user’s location and his communities. Note that a user can have multiple influential communities and an influential community can influence a user on multiple locations.

Figure 5b depicts the distribution of distances between users’ locations and their communities in New York and San Francisco in the two datasets. As we can see, most of the distances are short which indicates the communities are quite close to users’ locations. To illustrate that these

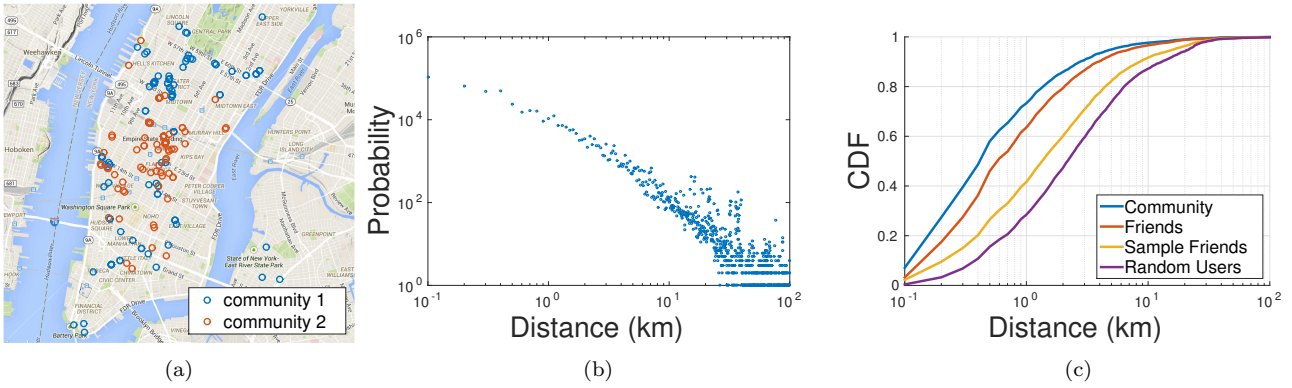


Figure 5: (a) A user’s two communities’ check-ins in Manhattan; (b) distribution of distances between users and their communities; (c) cumulative distribution function of distances between users and their communities, friends, sample friends and random users.

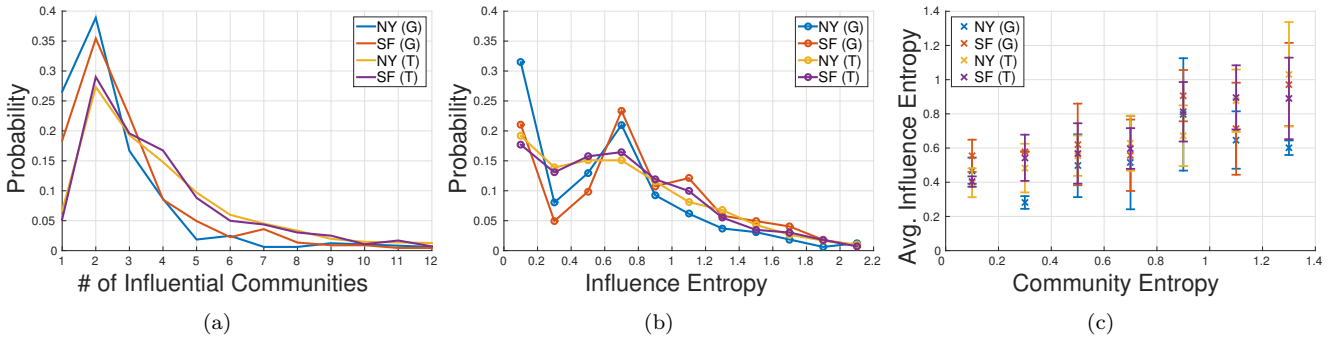


Figure 6: (a) Distribution of the number of influential communities; (b) distribution of influence entropies (bucketed by 0.2) (c) influence entropy vs. community entropy.

short distances are not due to the limits of the city areas, for each location of a user, we pick some random users in the city, summarize their frequent movement areas through clustering and find the minimal distance between their frequent movement areas and the location. In Figure 5c⁶, the curve of cumulative distribution function (CDF) for these random users (purple) is much lower than the one for communities (blue). This means that these random users are farther away from the users than communities. To show that community is a meaningful level to study mobility, we also calculate distances between a user and all his friends. The curve for friends (red) in Figure 5c is lower than the one for communities as well, meaning that a user is closer to his communities than to all his friends in general. As a user’s community is a subset of his friends, to illustrate that the shorter distances for communities than friends are not caused by frequent movement areas clustered from a small number of friends’ check-ins, for each community of a user, we randomly sample the same number of his friends to build a “virtual” community and calculate the distances between the user and his virtual communities. The CDF curve in Figure 5c (yellow) shows that these virtual communities are even farther away from users than all friends.

⁶The results in Figure 5c are based on the data from two cities in both datasets.

From the above analysis, we conclude that (1) communities have strong influences on users’ mobility and (2) community is a meaningful resolution to study users’ mobility.

4.2 Number of influential communities

Research shows that a user’s mobility is constrained geographically (see [9, 10]), e.g., a user normally travels in or around the city where he lives. Meanwhile, social relations are not restricted by geographic constraints. For instance, a user’s college friends as a community can spread all over the world. Now we focus on how many communities actually influence a user’s mobility i.e., how many influential communities a user has. Intuitively, this number should be small as each user only interacts with a limited number of communities in his daily life such as colleagues and family.

We plot the distribution of the number of user’s influential communities in Figure 6a. From two datasets, we can observe a similar result. Most of the users are influenced only by a small number of communities and there are more users who have two influential communities than others. For example, almost 30% of users in New York have two influential communities in the Twitter dataset.

Each location corresponds to an influential community. We proceed with studying how a user’s influential communities are distributed over his check-ins. We first propose a

notion named *influence entropy*, it is defined as

$$\text{infent}(u) = - \sum_{c \in C(u)} \frac{|CI(u, c)|}{|CI(u)|} \ln \frac{|CI(u, c)|}{|CI(u)|}$$

where $CI(u, c)$ represents u 's check-ins that are closest to the community c . The influence entropy is defined in the form of Shannon entropy: higher influence entropy indicates that the user's locations are close to his different communities more uniformly. Figure 6b depicts the distribution of users' influence entropies. As we can see, in New York (NY (T)), around 20% of users' influence entropies are between 0 and 0.2 which means they have one dominating influential community that is close to most of their locations. We also notice that there is a peak around 0.6 in all the cities. For example, if a user u 's 50% check-ins corresponds to one influential community and the other 50% corresponds to another one, then $\text{infent}(u) = 0.69$ which falls into this range. This shows that around 20% of users are influenced by their two major communities at a similar level.

Community entropy introduced in Section 3 is a notion for capturing a user's social diversity. We further study the relationship between community entropy and influence entropy. As shown in Figure 6c, more diverse a user's social relationship is, more probably his locations are distributed uniformly over his influential communities.

From the above analysis, we conclude that only a small number of communities have influences on users' mobility.

4.3 Communities under contexts

Influential communities are constrained by contexts. For instance, a user has lunch with his colleagues and spends time with his family near where he lives. Here, the lunch hour and the home location can be considered as social contexts, and the two communities (colleague and family) have impact on the user's behavior under each of the context, respectively. Thus it is interesting to study whether this hypothesis holds generally.

Temporal contexts. First, we focus on temporal contexts. The pair of contexts we choose are *Lunch* (11am–1pm) and *Dinner* (7pm–9pm) hours on Wednesday. For each user, we extract his check-ins during lunch and dinner time and find his influential communities w.r.t. these two contexts. We randomly choose four users and plot the distributions of their check-ins over their influential communities under these two contexts in Figure 7. As we can see, a user's communities behave quite differently on influencing his check-ins during lunch and dinner time. For example, the first user in New York in the Twitter dataset is only influenced by his community 3 during lunch time while communities 1 and 2 give him similar influences during dinner time. This simply reflects the fact that the people who users have lunch and dinner with are different. In addition, users' average influence entropies drop as well under different temporal contexts compared with the general case (see Table 3), this suggests that the influential communities tend to become more unique.

For each user during lunch (dinner) time, we create a vector where the i -th component counts the number of locations that are the closest to community i . We then exploit the cosine similarity between a user's lunch and dinner vectors as his *influence similarity*. The results are listed in Table 4. Note that, we also choose other pairs of temporal contexts

for analysis, such as working hours (9am–6pm) and nightlife (10pm–6am) and have similar observations.

Spatial contexts. Next we study the influence of spatial contexts. In each city, we pick two disjoint regions (called *Region 1* and *Region 2*, respectively) including Uptown and Downtown Manhattan in New York and Golden Gate Park and Berkeley in San Francisco. Then, we extract users' check-ins in these areas. By performing the same analysis as the one for temporal contexts, we observe similar results (see Figure 8, Table 3 and Table 4). Note that we choose the areas without special semantics in mind, e.g., business areas or residential areas.

| Influence entropy | NY (G) | SF (G) | NY (T) | SF (T) |
|-----------------------------|--------|--------|--------|--------|
| General | 0.56 | 0.73 | 0.69 | 0.70 |
| Temporal (<i>Lunch</i>) | 0.35 | 0.39 | 0.22 | 0.25 |
| Temporal (<i>Dinner</i>) | 0.27 | 0.43 | 0.30 | 0.31 |
| Spatial (<i>Region 1</i>) | 0.45 | 0.20 | 0.52 | 0.23 |
| Spatial (<i>Region 2</i>) | 0.42 | 0.21 | 0.61 | 0.26 |

Table 3: Influence entropy under different social contexts.

| Influence similarity | NY (G) | SF (G) | NY (T) | SF (T) |
|----------------------|--------|--------|--------|--------|
| Temporal | 0.80 | 0.74 | 0.67 | 0.66 |
| Spatial | 0.77 | 0.56 | 0.48 | 0.41 |

Table 4: Influence similarity w.r.t. social contexts.

From the above analysis, we can conclude that community impact is constrained under spatial and temporal contexts.

5. LOCATION PREDICTION

Location prediction can drive compelling applications including location recommendation and targeted advertising. On the other hand, it may also threat users' privacy [35]. Following the previous analysis, we continue to investigate whether it is possible to use community information to effectively predict users' locations, using machine learning techniques. More precisely, the question we want to answer is: given a user's community information, whether he will check in at a given place at a given time. Note that the time here is a certain hour on a certain day (Monday to Sunday).

We first list all the features in the community-based location prediction model. Then, we present the baseline predictors. Experimental results are described in the end.

5.1 Community-based location predictor

To predict whether a user will visit a certain location, we use one of his communities' information to establish the feature vector, i.e., the influential community of the location (see Section 4).

Community related features. Having chosen the community, we extract its following features for prediction.

- Distance between the community and the location. This is the distance between the location and the community's nearest frequent movement area.
- Community size. Number of users in the community.

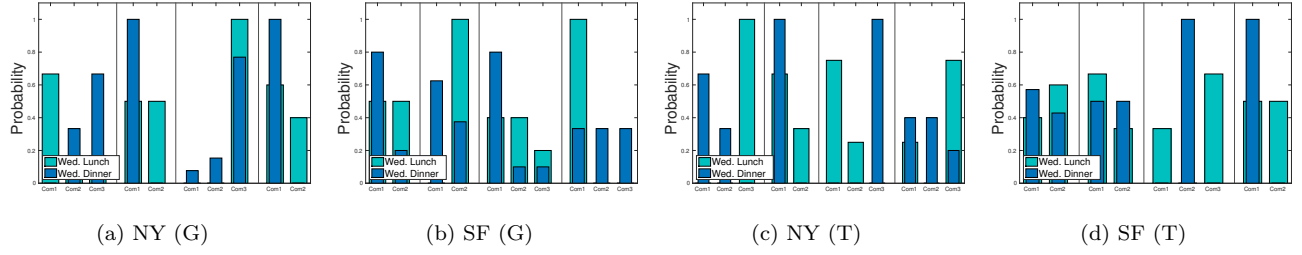


Figure 7: Distribution of influential communities on users' check-ins (temporal contexts).

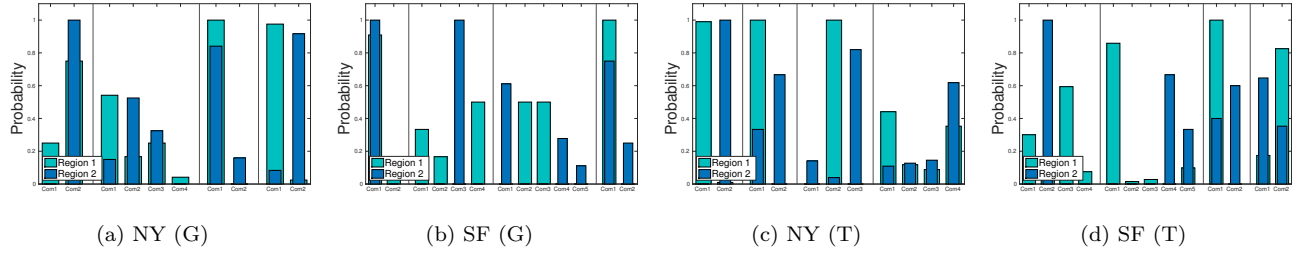


Figure 8: Distribution of influential communities on users' check-ins (spatial contexts).

- Number of the community's frequent movement areas.
- Community's total number of check-ins.
- Community connectivity. This is the ratio between the number of edges in the community and the maximal number of possible edges.

Time. Check-ins are related to time as well. Figure 9a (Figure 9b) plots the total number of check-ins in New York and San Francisco in a daily (weekly) scale. Since we aim to predict whether a user will check in at a place at a certain time, the time-related features we consider are the total number of check-ins at the time⁷ and the day (i.e., Monday to Sunday) from all users.

5.2 Baseline models

Sample friends. In our community-based predictor, each location corresponds to the user's nearest community. To illustrate the effectiveness of communities on predicting a user's mobility, in the first baseline model, for each location, we randomly sample the same number of friends as the community and use these friends to build a "virtual community" (as in Section 4). We then replace the community related features with this virtual community's corresponding ones. The time-related features of this model are exactly the same as the ones for the community-based model.

Friends. In the second baseline model, we consider a user's all friends instead of his communities. The features include the shortest distance from his friends to the location and the time-related features.

User. It has been shown in [10, 6] that a user's past mobility can predict his future mobility effectively. Therefore, we also extract features from a user himself to perform prediction. The features include the following.

⁷We consider time at a per hour unit, thus the feature is the number of check-ins of all the users at that hour.

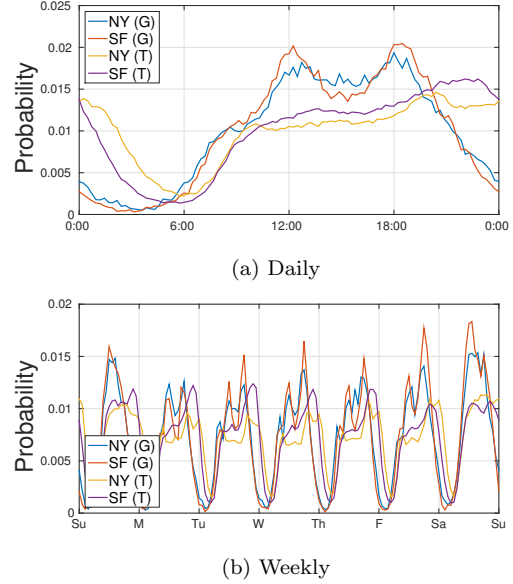


Figure 9: Check-in time in the datasets.

- The shortest distance from a user's frequent movement areas (through hierarchical clustering with cut-off distance equal to 500m) to the location.⁸
- The total number of check-ins during the day.
- The total number of check-ins during the hour.

User and community. In the last baseline model, we combine the features from the user's model and our community-based predictor.

⁸To avoid overfitting, we use half of each user's check-ins to discover his frequent movement areas and the other half are used for training and testing the model.

5.3 Metrics

We partition the cities into 0.001×0.001 degree latitude and longitude cells, a user is said to be in a cell if he has been to any place belonging to the cell. Let TP , FP , FN and TN denote true positives, false positives, false negatives and true negatives, respectively. The metrics we adopt for evaluation include (1) Accuracy,

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |FN| + |TN|};$$

(2) F1 score,

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}, \text{ with}$$

$$Precision = \frac{|TP|}{|TP| + |FP|}, \quad Recall = \frac{|TP|}{|TP| + |FN|};$$

and (3) AUC (area under the ROC curve).

5.4 Experiment setup

We build a classifier for each user. A classifier needs both positive and negative examples. So far we only have the positive ones, i.e., a user visits a location. To construct the negative examples, for each location a user visits, we randomly sample a different location (within the city) as the place that he does not visit at that moment. In this way, a balanced dataset for each user is naturally formed. As in the data analysis, we only focus on active users who have at least 100 check-ins in the city. For each user, we sort his check-ins chronologically and put his first 80% check-ins for training the model and the rest 20% for testing. The machine learning classifier we exploit here is logistic regression. In all sets, we perform 10-fold cross validation.

5.5 Results

Performance in general. As depicted in Figure 10, our community-based predictor’s performance is promising and it outperforms two baseline models that exploit friends’ information. Especially for the sample friends model, the community-based model is almost 20% better among all three metrics in the Twitter dataset. By studying logistic model’s coefficients, the most important feature is the distance between the community and the location, followed by the community connectivity and size.

On the other hand, two predictions that are based on user’s own information perform better than our community-based predictor. Also, the predictor combining user and community information does not improve the performance. This indicates that a user’s past check-ins are the most useful information for predicting where he will be in the future which also validates the results proposed in [10, 6].

Prediction vs. community entropy. In Figure 11, we bucket community entropy by intervals of 0.2 and plot its relationship with the prediction results (AUC). As we can see, with the increase of community entropy, the AUC grows for the community-based model which means the predictor works better for users with high community entropies. For example, the AUC value increases more than 5% in San Francisco in the Gowalla dataset (community entropy from $[0, 0.2)$ to $[1.2, 1.4)$).

We further calculate the Pearson’s correlation coefficient⁹ between community entropy and our prediction results. In the Twitter dataset, the correlation coefficient for New York and San Francisco is 0.88 and 0.97 respectively,¹⁰ indicating that community entropy and the prediction results are strongly correlated. This validates our intuition that a user with high social diversity is clearly influenced by his communities. We can conclude that community information can be explored to achieve promising location predictions, especially for those users with high community entropies.

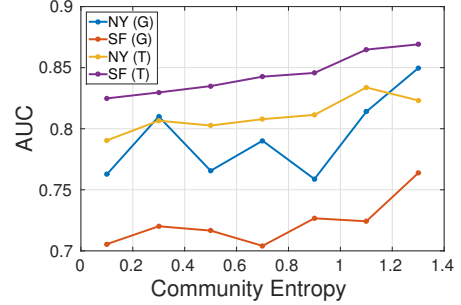


Figure 11: AUC as a function of community entropy, values of the Pearson’s correlation: 0.60 (NY (G)), 0.75 (SF (G)), 0.88 (NY (T)), 0.97(SF (T)).

Difference between cities. In Figure 10 and Figure 11, we observe that the prediction results are different between two cities. New York has the better performance than San Francisco in the Gowalla dataset. On the other hand, the prediction results are similar in the Twitter dataset. The reason for different performances in different cities could be due to the density of the cities (e.g., New York’s population density is higher than San Francisco), or the adoption of LBSN services by users in different cities. We leave the investigation as a future work.

5.6 Other strategies to choose communities

So far, we have shown that exploring community information can lead to effective location prediction. The community we choose is the one that has the closest frequent movement area to the target location. We would like to know if other strategies to choose community can achieve similar results. We consider three strategies including choosing the community with most users (**max-size**), the community with highest connectivity (**max-con**) and random community (**random**). Table 5 summarizes the prediction performances in New York in the Twitter dataset. As we can see, our original strategy outperforms these three. Among these three strategies, **max-con** performs slightly better than the other two, but it is still relatively worse than our original strategy to choose community. This again validates our observation in Section 4 that influential communities are constrained by contexts (spatially or temporally), in other words one community cannot influence every location of the user.

⁹Pearson’s correlation coefficient is the covariance of two variables divided by the product of their standard deviations.

¹⁰The two values are slightly smaller for the Gowalla dataset, which is probably due to the fact that the Twitter dataset contains more information on social relations than the Gowalla dataset (see discussions in Section 3).

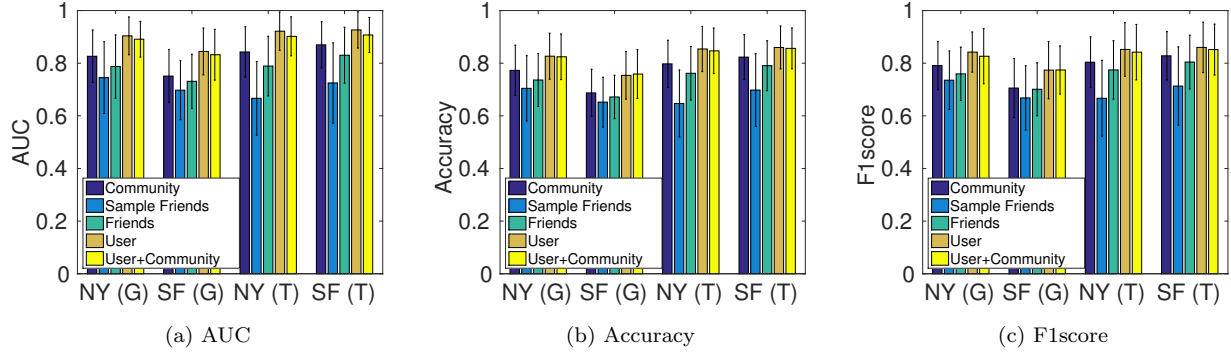


Figure 10: Prediction results.

| | AUC | Accuracy | F1score |
|-----------|------|----------|---------|
| Community | 0.83 | 0.78 | 0.79 |
| max-size | 0.73 | 0.72 | 0.74 |
| max-con | 0.74 | 0.73 | 0.74 |
| random | 0.71 | 0.71 | 0.72 |

Table 5: Performance of community-choosing strategies.

5.7 Comparison with the PSMM model

In [10], the authors establish a mobility model (PSMM) for each user based on his past check-ins. The assumption behind this model is that a user’s mobility is mainly centered around two states such as home and work. Each state is modeled as a bivariate Gaussian distribution and the total mobility is then formalized into a dynamic Gaussian mixture model with time as an independent factor. The check-ins that do not fit well with the two states are considered as social check-ins and are modeled through another friends-based distribution. We implement the PSMM model and compare its performance with our community-based predictor. Each user’s first 80% check-ins are used for training his PSMM model. For testing, besides the rest 20% check-ins, we also construct the same number of locations that the user does not go at the moment (as our classification setup). As the PSMM model’s output is the exact location of the user, we consider the prediction is correct when the output location is within 1km of the real location. Table 6 shows the accuracy between our model and PSMM. In all the datasets, our community-based predictor significantly outperforms PSMM. As suggested in [33], this is probably because two states are not enough to capture a user’s mobility in a city. Moreover, a user’s check-in data is also too sparse to train a good PSMM model. We leave the further investigation as a future work.

| | NY (G) | LA (G) | NY (T) | SF (T) |
|-----------|--------|--------|--------|--------|
| Community | 0.76 | 0.67 | 0.78 | 0.81 |
| PSMM | 0.55 | 0.60 | 0.67 | 0.65 |

Table 6: Comparison with PSMM on prediction accuracy.

6. RELATED WORK

Thanks to the emerging of LBSNs, mobility as well as its connection with social relations have been intensively stud-

ied [9, 34, 15]. There are mainly two directions of research going on in the area. One direction is to use the location information from LBSNs to predict friendships (see e.g. [19, 13, 12, 6, 33, 28, 40]), the other studies the impact from friendships on locations [1, 6, 10, 33, 24] which is what we focus on in the current work.

Backstrom, Sun and Marlow [1] study the friendship and location using the Facebook data with user-specified home addresses. They find out that the friendship probability as a function of home distances follows a power law, i.e., most of friends tend to live closely. They also build a model to predict users’ home location based on their friends’ home. Their model outperforms the predictor based on IP addresses. The authors of [6] use the Facebook place data to study check-in behaviors and friendships. They train a logistic model to predict users’ locations. Besides that, they also investigate how users respond to their friends’ check-in and use the location data to predict friendships. Cho, Myers and Leskovec [10] investigate the mobility patterns based on the location data from Gowalla, Brightkite as well as data from a cellphone company. Based on their observation, they build a dynamic Gaussian mixture model for human mobility involving temporal, spatial and social relations features. Sadilek, Kautz and Bigham [33] propose a system for both location and friendship prediction. For location prediction, they use dynamic Bayesian networks to model friends’ locations (unsupervised case) and predict a sequence of locations of users over a given period of time. McGee, Caverlee and Cheng [24] introduce the notion of social strength based on their observation from the geo-tagged Twitter data and incorporate it into the model to predict users’ home locations. Experimental results show that their model outperforms the one of [1]. Jurgens in [17] proposes a spatial label propagation algorithm to infer a user’s location based on a small number initial friends’ locations. Techniques such as exploiting information from multiple social network platforms are integrated into the algorithm to further improve the prediction accuracy.

The main difference between previous works and ours is the way of treating friends. We consider users’ friends at a community level while most of them treat them the same (except for the paper [24] which introduce ‘social strength’, which is based on common features but not on communities). Moreover, our location predictor doesn’t need any user’s own information but his friends’ to achieve a promising result, especially for users’ with high community entropies. Other mi-

nor differences include the prediction target: we want to predict users' certain locations in the future not their home [1, 24, 17] or a dynamic sequences of locations [33].

We focus on understanding users' mobility behavior from social network communities. The authors of [4] tackle the inverse problem, i.e., they exploit users' mobility information to detect communities. They first attach weights to the edges in a social network based on the check-in information, then the social network is modified by removing all edges with small weights. In the end, a community detection algorithm (louvain method[2]) is used on the modified social graph to discover communities. The experimental results show that their method is able to discover more meaningful communities, such as place-focused communities, compared to the standard community detection algorithm.

More recently, Brown et al. [3] analyze mobility behaviors of pairs of friends and groups of friends (communities). They focus on comparing the difference between individual mobility and group mobility. For example, they discover that a user is more likely to meet a friend at a place where they have not visited before; while he will choose a familiar place when meeting a group of friends.

7. CONCLUSION AND FUTURE WORK

In this paper, we have studied the community impact on user's mobility. Analysis leads us to several important conclusions: (1) communities have a stronger impact on users' mobility; (2) each user is only influenced by a small number of communities; and (3) different communities have influences on mobility under different spatial and temporal contexts. Based on these, we use machine learning techniques to predict users' future locations focusing on community information. The experimental results on two types of real-life social network datasets are consistent with our analysis and show that our prediction model is very effective. The scripts for conducting the analysis and experiments as well as the Twitter dataset are available upon request.¹¹

In the future, we plan to extend our work in several directions. First, we have shown in this paper that communities can be exploited to achieve a promising location prediction. We are also interested in extending our work to other applications such as location recommendation. It is possible to redesign the cost function in matrix factorization based methods for location recommendation by taking into account community information. Second, we would like to conduct the analysis of community impact on other social behaviors such as information sharing or interests adoption. Third, in a broader point of view, our current work is actually a demonstration of the communities' effect on human behaviors. As pointed by [39], community is the most meaningful resolution to study social network. Therefore, we also plan to investigate a user's role in his social network based on the structure of his communities.

8. REFERENCES

- [1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. 19th International Conference on World Wide Web (WWW)*, pages 61–70. ACM, 2010.
- [2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [3] C. Brown, N. Lathia, C. Mascolo, A. Noulas, and V. Blondel. Group colocation behavior in technological social networks. *PLoS ONE*, 9(8):e105816, 2014.
- [4] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo. The importance of being placefriends: discovering location-focused online communities. In *Proc. ACM Workshop on Online Social Networks (WOSN)*, pages 31–36. ACM, 2012.
- [5] S. Castles, M. J. Miller, and G. Ammendola. *The Age of Migration: International Population Movements in the Modern World*. Taylor & Francis, 2005.
- [6] J. Chang and E. Sun. Location³: How users share and respond to location-based data on social networking sites. In *Proc. 5th AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 74–80. The AAAI Press, 2011.
- [7] X. Chen, J. Pang, and R. Xue. Constructing and comparing user mobility profiles for location-based services. In *Proc. 28th ACM Symposium on Applied Computing (SAC)*, pages 261–266. ACM, 2013.
- [8] X. Chen, J. Pang, and R. Xue. Constructing and comparing user mobility profiles. *ACM Transactions on the Web*, 8(4):article 21, 2014.
- [9] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *Proc. 5th AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 81–88. The AAAI Press, 2011.
- [10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. 17th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1082–1090. ACM, 2011.
- [11] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.
- [12] D. J. Crandalla, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.
- [13] J. Cranshaw, E. Toch, J. Hone, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proc. 12th ACM International Conference on Ubiquitous Computing (UbiComp)*, pages 119–128. ACM, 2010.
- [14] H. Gao, J. Tang, X. Hu, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proc. 7th ACM Conference on Recommender Systems (RecSys)*, pages 93–100. ACM, 2013.
- [15] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proc. 6th AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 114–121. The AAAI Press, 2012.

¹¹ Preliminary results of this work are reported as a poster [27].

- [16] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [17] D. Jurgens. That’s what friends are for: Inferring location in online social media platforms based on social relationships. In *Proc. 7th AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2013.
- [18] A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *CoRR*, abs/0908.1062, 2010.
- [19] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proc. 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*, page 34. ACM, 2008.
- [20] B. Liu and H. Xiong. Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proc. 13th SIAM International Conference on Data Mining (SDM)*, pages 396–404. SIAM, 2013.
- [21] E. L. Martelot and C. Hankin. Fast multi-scale detection of relevant communities in large-scale networks. *The Computer Journal*, 56(9):1136–1150, 2013.
- [22] J. J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Proc. 26th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 548–556. NIPS, 2012.
- [23] J. J. McAuley and J. Leskovec. Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data*, 8(1):article 4, 2014.
- [24] J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *Proc. 22nd ACM International Conference on Information & Knowledge Management (CIKM)*, pages 459–468. ACM, 2013.
- [25] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [26] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn. Community-enhanced de-anonymization of online social networks. In *Proc. 21st ACM Conference on Computer and Communications Security (CCS)*, pages 537–548. ACM, 2014.
- [27] J. Pang and Y. Zhang. Exploring communities for effective location prediction (poster paper). In *Proc. 24th World Wide Web Conference (Companion Volume) (WWW)*, pages 87–88. ACM, 2015.
- [28] H. Pham, C. Shahabi, and Y. Liu. EBM: an entropy-based model to infer social strength from spatiotemporal data. In *Proc. 2013 ACM International Conference on Management of Data (SIGMOD)*, pages 265–276. ACM, 2013.
- [29] D. Quercia, R. Schifanella, L. M. Aiello, and K. McLean. Smelly maps: the digital life of urban smellscape. In *Proc. 9th AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 237–236. The AAAI Press, 2015.
- [30] A. Rényi. On measures of information and entropy. In *Proc. 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960.
- [31] M. Rosvall and C. T. Bergstrom. Maps of information flow reveal community structure in complex networks. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
- [32] M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS ONE*, 6(4):e18209, 2011.
- [33] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proc. 5th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 459–468. ACM, 2012.
- [34] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *Proc. 5th AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 329–336. The AAAI Press, 2011.
- [35] R. Shokri, G. Theodorakopoulos, J.-Y. L. Boudec, and J.-P. Hubaux. Quantifying location privacy. In *Proc. 32nd IEEE Symposium on Security and Privacy (S&P)*. IEEE CS, 2011.
- [36] K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks (extended abstract). In *Proc. 16th International Conference on World Wide Web (WWW)*, pages 1275–1276. ACM, 2007.
- [37] J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proc. 6th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 587–596. ACM, 2013.
- [38] J. Yang, J. J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *Proc. 13th IEEE International Conference on Data Mining (ICDM)*, pages 1151–1156. IEEE CS, 2013.
- [39] J. Yang, J. J. McAuley, and J. Leskovec. Detecting cohesive and 2-mode communities in directed and undirected networks. In *Proc. 7th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 323–332. ACM, 2014.
- [40] Y. Zhang and J. Pang. Distance and friendship: A distance-based model for link prediction in social networks. In *Proc. 17th Asia-Pacific Web Conference (APWeb)*, LNCS. Springer, 2015. To appear.
- [41] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *Proc. 19th International Conference on World Wide Web (WWW)*, pages 1029–1038. ACM, 2010.
- [42] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: when urban air quality inference meets big data. In *Proc. 19th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1436–1444. ACM, 2013.
- [43] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, 5(1), 2011.
- [44] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *Proc. 18th International Conference on World Wide Web (WWW)*, pages 791–800. ACM, 2009.